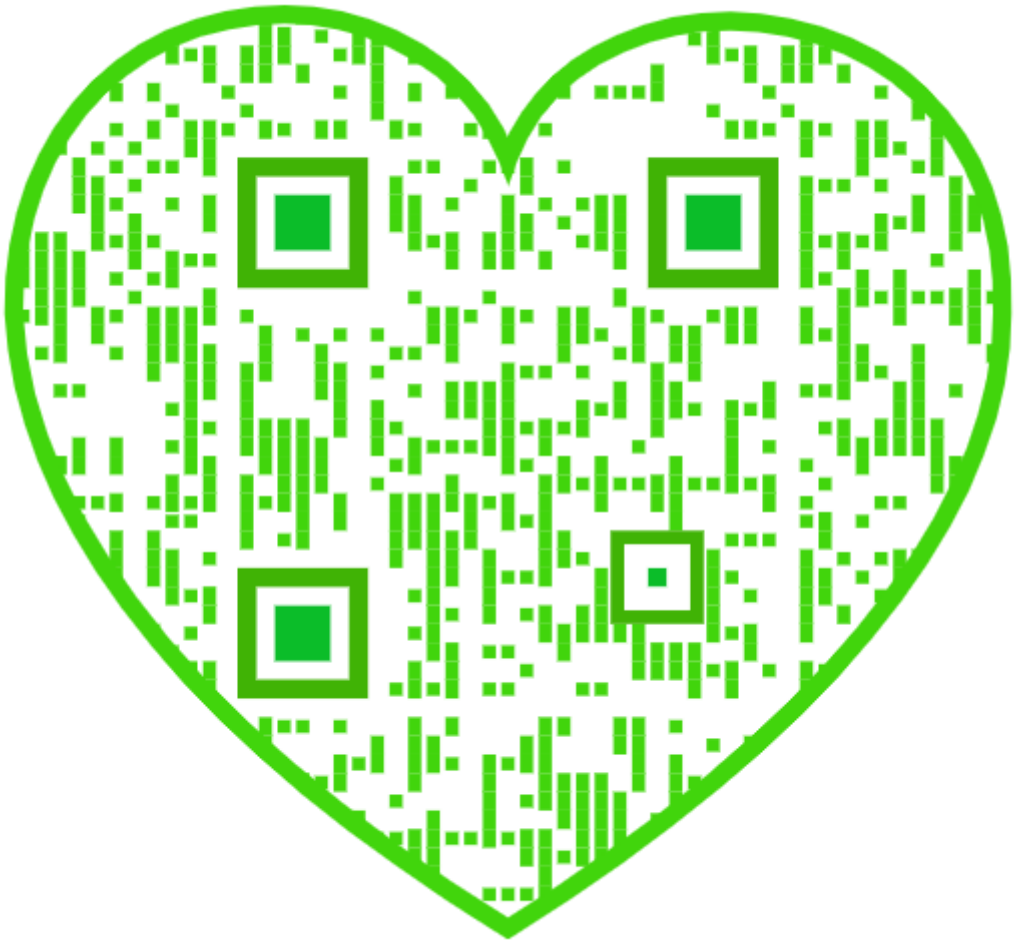
Master in Artificial Intelligence

Deployment II



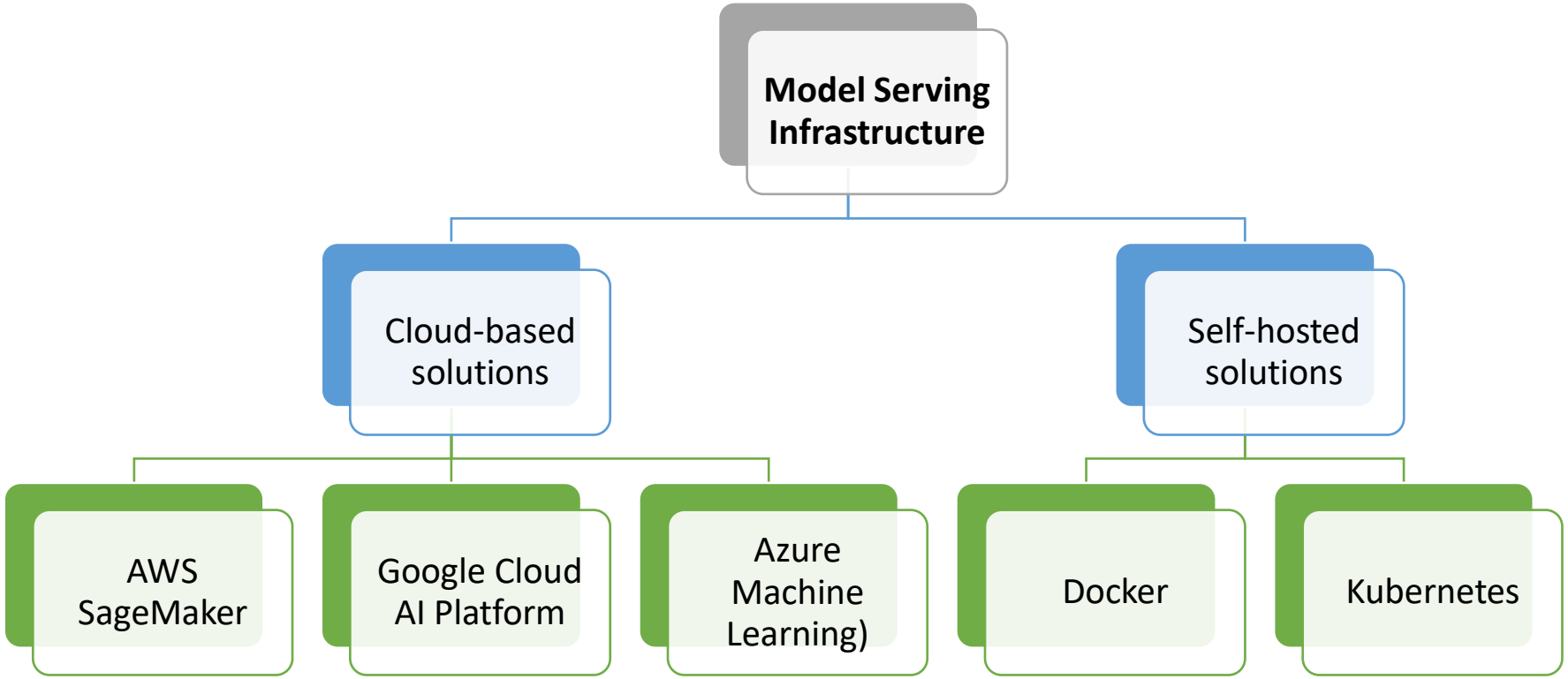
Purpose

The purpose of the section is to help you learn how to deploy trained models into production environments to become a Successful Artificial Intelligence (AI) Engineer

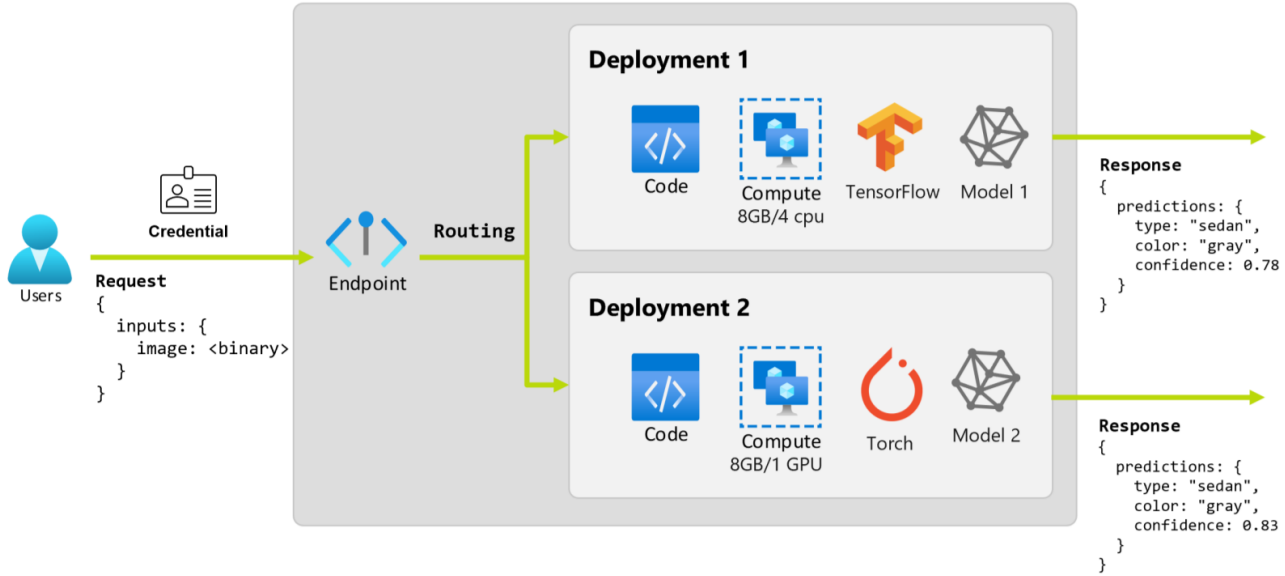
At the end of this lecture, you will learn the following

- **How to deploy trained models into production environments, ensuring they integrate smoothly with existing systems and meet performance requirements**

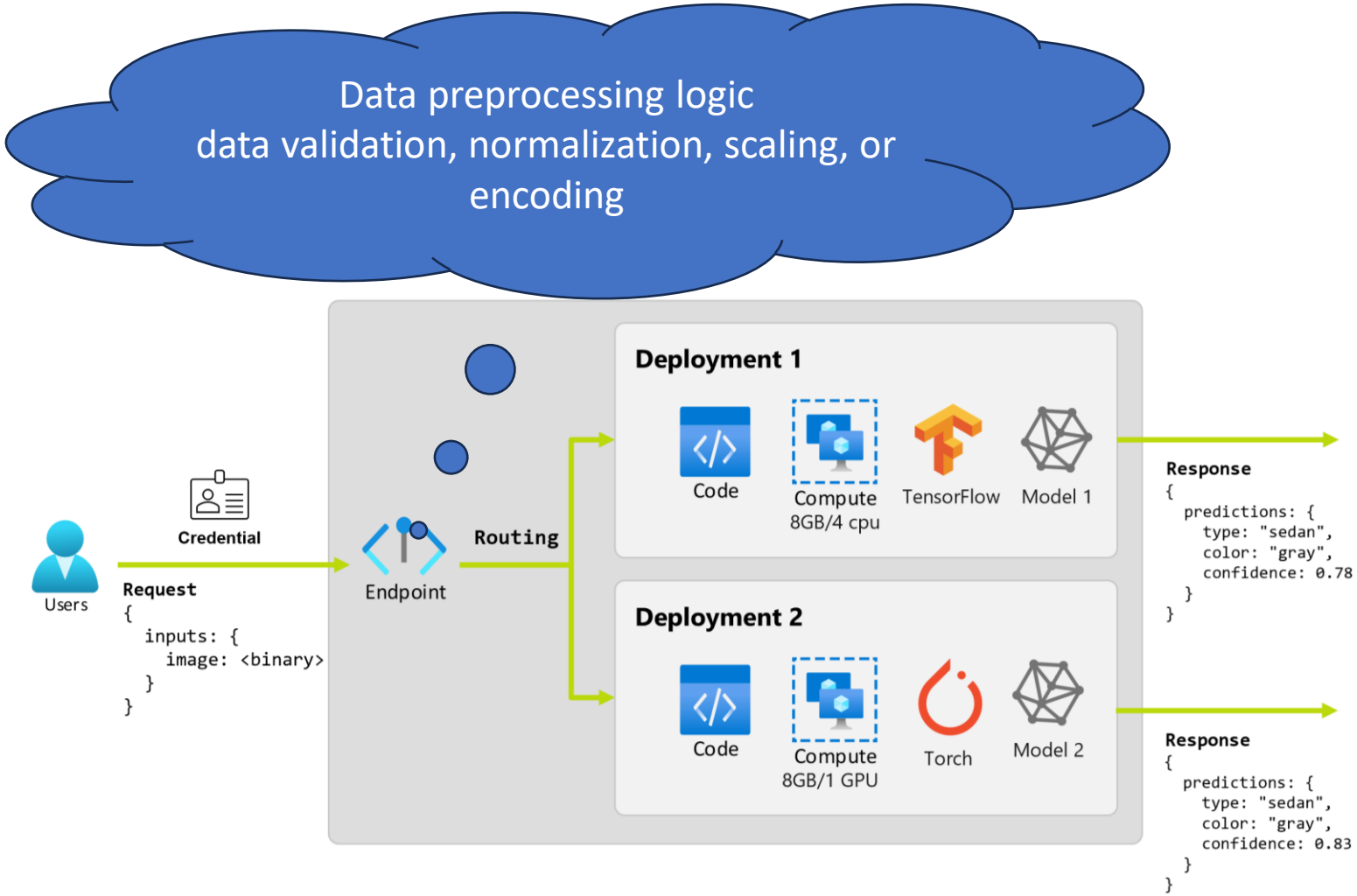
Model Serving Infrastructure



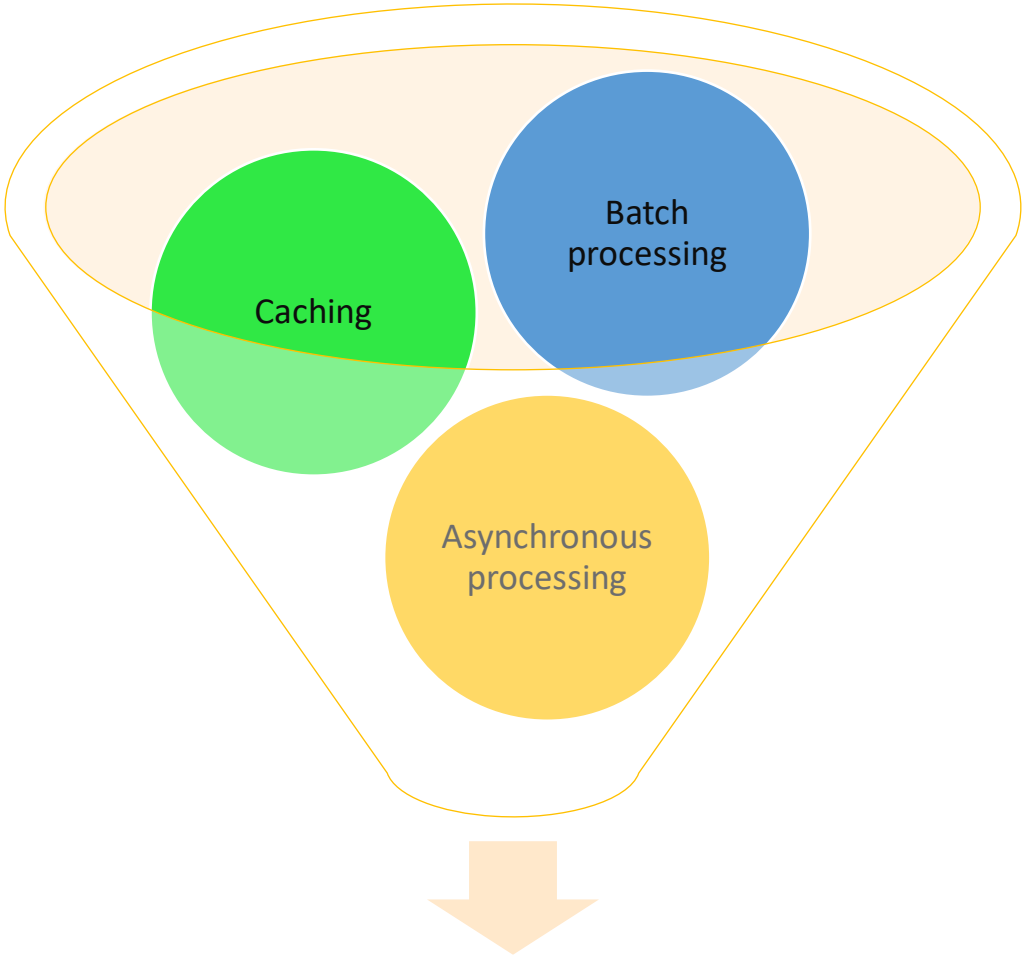
API Endpoint Creation



Input Data Preprocessing



Performance Optimization



Improve response times and scalability

Monitoring and Logging

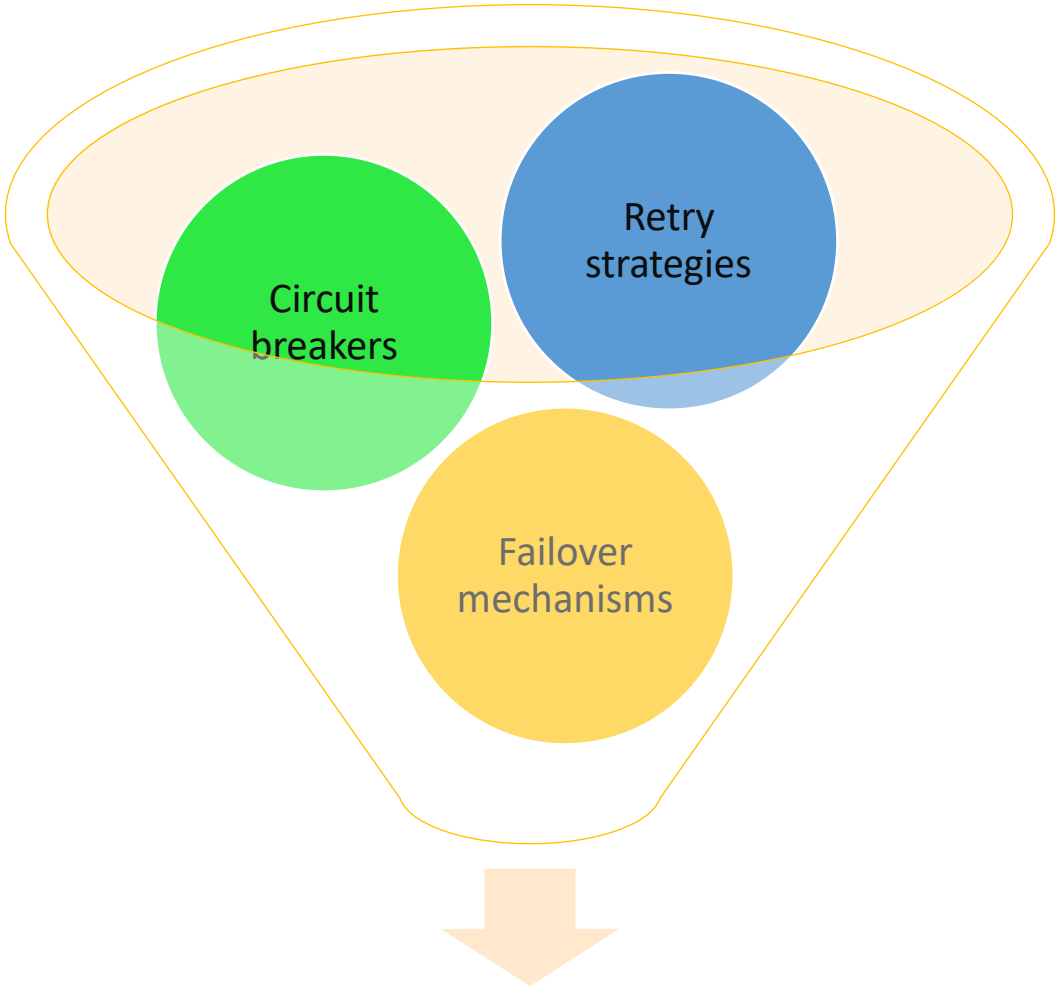
Track

- Performance
- Usage
- Health of the deployed models in real-time

Monitor key metrics

- Response time
- Throughput
- Error rates
- Resource utilization

Error Handling and Resilience



System resilience and reliability

Security and Authorization

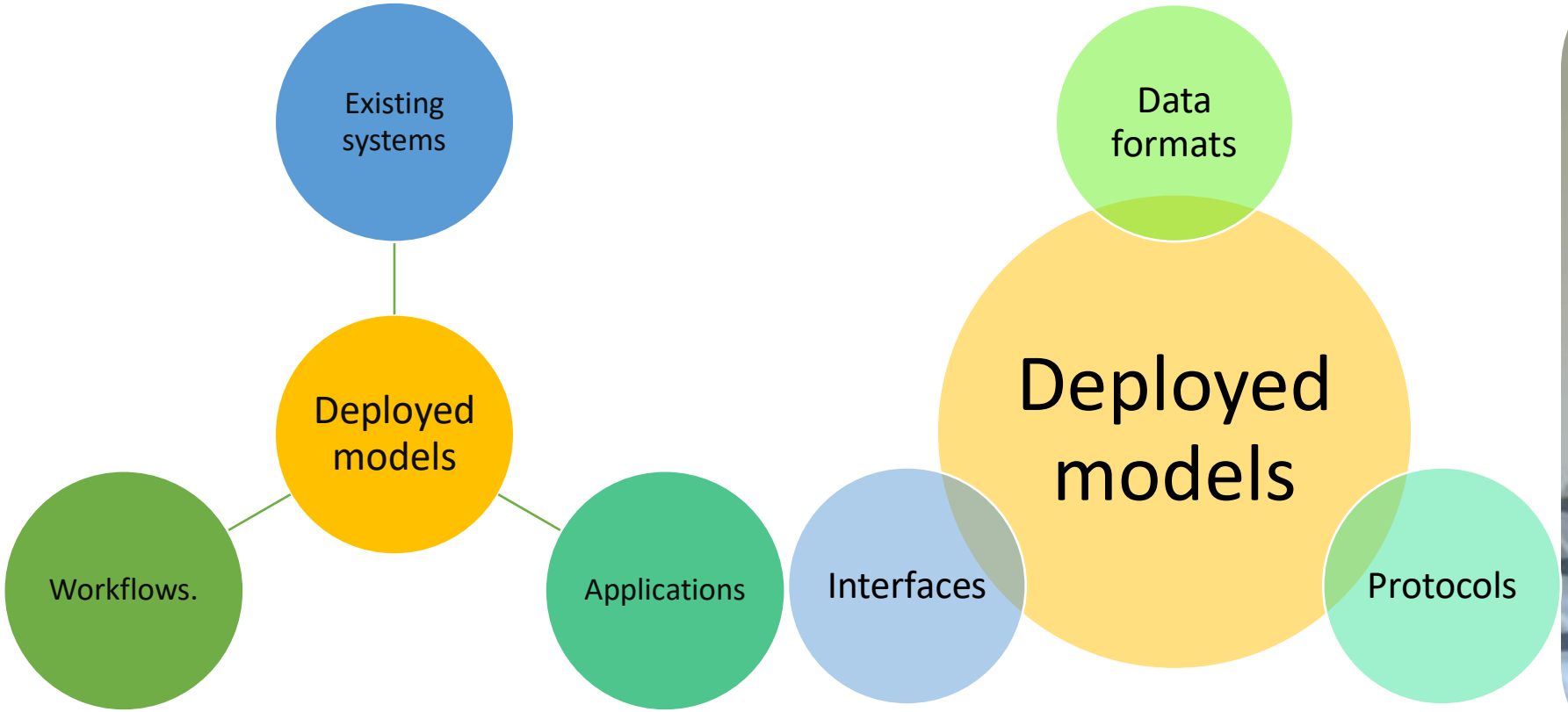
Secure the deployed models

- Authentication
- Authorization
- Encryption mechanisms

Ensure compliance

- Data privacy regulations (e.g., GDPR, HIPAA)
- Industry security standards

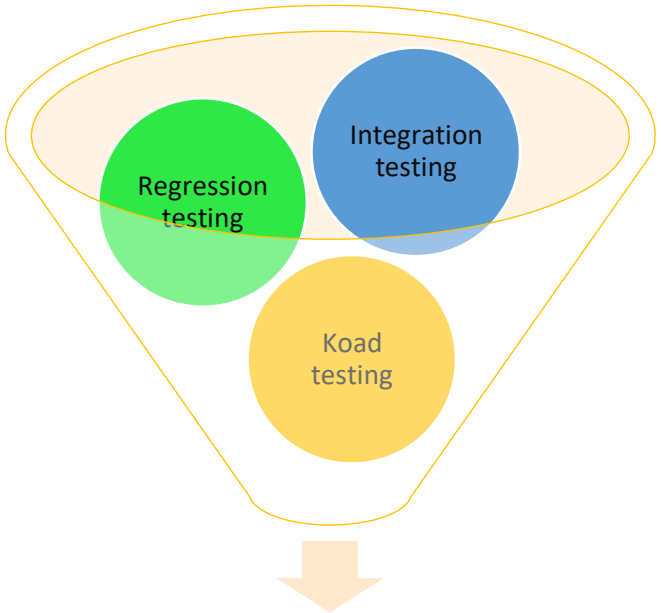
Integration with Existing Systems



Testing and Validation

Conduct thorough testing and validation of the deployed models

- Correctness
- Performance
- Reliability under various conditions



Documentation and Training

Document

- Deployment process
- API endpoints
- Input/output formats
- Usage guidelines for developers and users

Provide

- Training and support
- Relevant stakeholders
- How to interact with
- Use the deployed models effectively

How to deploy trained models into production environments

Trained Model

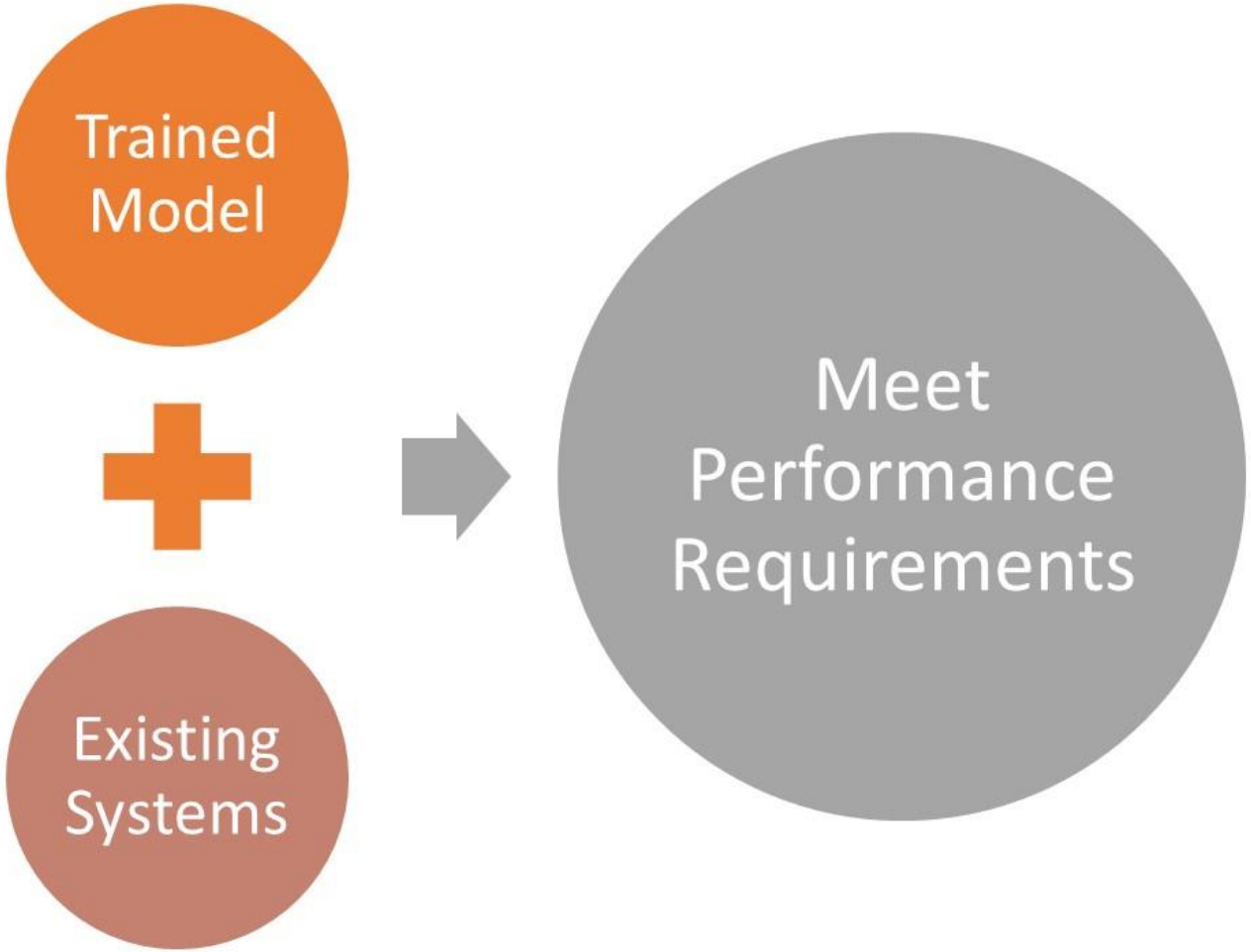
+

Existing Systems

Meet Performance Requirements

What is next?

Example of How to deploy trained models into production environments



Master in Artificial Intelligence

Deployment II